计算模型在道德认知研究中的应用

张银花 李红 吴寅*

（深圳大学师范学院心理学院；深圳市情绪与社会认知科学重点实验室，深圳 518060）

摘 要 道德认知关注道德心理背后的信息加工。近年来，研究者开始将计算模型应用于道德认知研究，以探索道德认知如何在大脑中实现。但目前研究者对道德认知进行计算建模的研究处于起步阶段。计算模型（漂移扩散模型、效用模型、强化学习模型和分层高斯过筛器模型）在道德认知行为和生理研究上的运用量化了道德决策、道德判断和道德推理背后的认知过程和神经机制。此外，这一新进展对理解反社会行为和精神障碍等有所助益。最后，计算建模有待完善，未来研究需要关注其潜在的问题。

关键词 道德认知；计算模型；道德决策；道德判断；道德推理

近日，贺建奎团队完成了"首例基因编辑婴儿诞生"的实验（参见，《参考消息》2019–01–23）。对此，许多人表示，贺建奎的行为明显违反了伦理。而且，贺建奎的道德品质也受到质疑。这一事件的争论焦点在于伦理道德问题。目前，研究者对道德认知领域进行了大量研究，但尚未阐明解决道德问题特有的认知机制。随着对行为数据的计算建模方法日臻成熟，研究者已开始将计算模型运用于道德认知领域。计算模型以数学函数的形式定量地描述选项特征（如代价、收益和等待时间）如何转换为效价，进而影响决策（Brown, 2014; Charpentier & O'Doherty, 2018; Konovalov, Hu, & Ruff, 2018）。最近的研究已经使用这种方法描述道德效价的计算，即道德问题的外部特征（如利益、伤害等）如何转化为内部效用，以及该效用如何指导道德决策、判断和推理（Hackel & Zaki, 2018; Hutcherson, Bushong, & Rangel, 2015; Siegel, Estrada, Crockett, & Baskin-Sommers, 2019; Siegel, Mathys, Rutledge, & Crockett, 2018; Yu, Siegel, & Crockett, 2019）。本文将回顾道德认知的内涵、计算模型在道德认知领域的运用以及其如何促进我们对道德认知过程和相关神经机制的理解。

1 道德认知

贺建奎事件包括了（1）贺建奎做出基因编辑婴儿的决策（decision-making）；（2）读者对其特定选择是否符合道德做出判断（judgment）；（3）进一步，读者会对其的道德品质做出推理（inference）。以上对应了道德认知的三个维度——道德决策、判断和推理（本文对道德认知的分类参照了 Yu 等人（2019）的划分方式†，参见 Yu et al., 2018）。它们的定义如下：道德决策是指人们做出影响他人利益的选择；道德判断是指人们判断行为或心理状态（如情绪、态度等）是否符合道德的过程，有时包含对某种行为是否应被惩罚或奖励的判

* 收稿日期：2019-04-22

本文系国家自然科学基金(31872784, 31600923)，广东省教育厅教育科学规划青年项目（2018GXJK150），深圳大学新教师科研启动项目的研究成果之一。

通信作者：吴寅， E-mail: yinwu0407@gmail.com

† Yu 等人（2019）将道德认知分为道德决策、道德判断和道德推理三个维度。本文参考了这种划分方式，并在其基础上对道德认知阐述时进行了补充和扩展。此外，Yu 等人提出一个以伤害厌恶为核心的统一计算框架来解释道德认知，而本文综合考虑了不同模型（漂移扩散模型、效用模型、强化学习模型、分层高斯过筛器和效用模型）在道德认知研究中的应用。

断；道德推理是人们基于对道德相关行为的观察而形成对行为者道德品质（如善或恶）的信念（Yu et al., 2019）。以下我们将从这三个维度展开对道德认知的心理学研究的介绍。

1.1 道德决策

道德决策涉及个体的选择是否损害他人的利益。人有自利倾向（Gray, 1987），会在诚实/不诚实、公平/不公平和慷慨/自私等决策之间进行权衡。以诚实决策为例，人们会做出诚实决策（放弃由不诚实带来的额外收益），还是不诚实决策（获得额外收益）？以往研究指出，相比诚实个体，不诚实个体放弃由不诚实决策获取的利益的时间更长（Greene & Paxton, 2009）。这表明不诚实个体在放弃不诚实利益的过程可能产生更多的认知需求。而且当人们做不诚实选择时，其心理和生理上均会感到不适（Cohn, Fehr, & Maréchal, 2014; Gächter & Schulz, 2016; Gamer, Rill , Vossel, & Gödert, 2006）。为了减轻这种不适感，个体会减少不道德行为。此外，背外侧前额叶皮层损伤的个体对诚实问题的敏感性降低（Zhu et al., 2014），杏仁核的激活程度与个体不诚实行为的历史呈负相关——个体在当前不诚实决策中杏仁核激活的降低程度预示着下一决策中不诚实的增加程度（Engelmann & Fehr, 2016; Garrett, Lazzaro, Ariely, & Sharot, 2016）。这表明背外侧前额叶皮层和杏仁核对诚实决策的重要作用。综上，决策往往需要在物质利益和道德价值之间权衡，但当选择道德决策时，对物质利益的权重会减小，人们更加关心如诚实、慷慨等道德价值。

1.2 道德判断

道德判断基于道德决策，指人们判断决策或决策者应被给予奖励还是施加惩罚。电车困境是研究道德判断的常用范式——想象一辆失控的电车即将撞死铁轨上的五名工人，决策者可以选择什么都不做，五名工人会死亡；或扳动开关将电车转向一个侧道，那里的一名工人会死亡（Kamm, 2015）。根据人们对两种选择的道德认可程度，Greene（2007）提出道德判断的双过程模型——义务性和功利性道德判断，即支持决策者什么都不做是一种义务性判断（在义务论道德体系下，“不可主动杀人”是一项道德义务），而支持决策者牺牲一个人拯救五个人是一种功利性判断（在功利主义道德体系下，一人死亡比五人死亡价值更高）；前者由情感驱动，是快速、自动的过程；后者由认知驱动，是缓慢、需要动机和认知资源参与的过程。研究表明，在产生共情的情况下，个体做出义务性道德判断的频率增加；而个体与受害者接触较少或倾向于理性思维方式时，做出功利性道德判断的频率增加（Elqayam, Wilkinson, Thompson, Over, & Evans, 2017; Greene, 2014）。进一步发现，血清素通过增加个体对伤害他人的厌恶，降低人们做出功利性判断的可能性（Crockett, Clark, Hauser, &Robbins, 2010）。相反，腹内侧前额叶皮层损伤的个体做出异常高的功利性判断（Koenigs et al., 2007），表明腹内侧前额叶皮层是直觉的、情感系统的关键神经基质，对正常的道德判断至关重要。综上，伤害厌恶是一种亲社会情绪，直接影响道德判断和道德行为，也在治疗反社会和攻击性行为中的应用有一定的启示。

1.3 道德推理

道德推理的核心是由可观察的、已知的现象（如他人的外显行为）推断内隐的、未知的状态（如他人行为背后的动机或他人的道德品质）。近年来，道德推理研究的焦点是对行为的评价，即个体指出影响他们进行道德推理的特征。研究表明，负性行为（如偷盗）比正性行为（如捐赠）更能代表个体的道德品质（Eisenegger, Naef, Snozzi, Heinrichs, & Fehr, 2010; Uhlmann, Pizarro, & Diermeier, 2015）。捐赠可能由其他动机驱动（如维护自己的社会地

位），供人推理的信息比较少；而偷盗的动机大多是负面的（如利己、反社会等），从而更容易推断偷盗者的道德品质。这表明个体进行道德推理时受信息量高低的影响。另有研究表明，人们通常给予伪君子（一边谴责不道德行为，一边做着不道德行为的人）负性评价（Jordan, Sommers, Bloom, & Rand, 2017; Levine, Barasch, Rand, Berman, & Small, 2018）。然而，伪君子通过承认不道德行为来避免向他人发出虚假信号，人们对他们的评价则没那么负性。这表明人们对行为者发出虚假的道德信号比较反感。此外，有害的行为（如从超市偷了一只死鸡）比无害但不洁的行为（如煮食自己死去的宠物狗）更不道德，但后者中行为者的道德品质更低下（Uhlmann & Zhu, 2014）。这表明，以个体品质为中心的道德推理，通常比行为的后果或是否违背道德准则更重要。综上，道德推理是深思熟虑的和直觉的过程（Garon, Lavallée, Estay, & Beauchamp, 2018）。

2 计算模型

计算机的发展与应用加快了计算建模研究的速度，为科学研究提供了更先进、严谨的手段。计算模型以数学函数的形式，将实验中可观察到的变量（如刺激、结果或过去的经验）与近期的行为联系起来，并对行为产生的不同算法假设进行量化。研究者们通过将实验数据与模型进行拟合，探究行为背后的算法，使用精确的数学模型更好地理解行为数据。

近年来，计算模型在心理学研究领域被广泛应用，如感知觉、决策、记忆和学习等方面。Jiang, Summerfield 和 Egner（2016）将计算模型与行为和神经成像数据结合起来，揭示了视觉对象不同的特征预期（和注意力）如何在驱动感知决策和神经表征的过程中相互作用，并表明视觉对象是预测视觉的选择单位。简单地说，当视觉对象的一个特征在预期之外时，这种预测误差会传播到其他特征，使该对象的其他特征也在预期之外，于是该视觉对象整体在预期之外。此外，人们也会从经验中获得的价值预期生成决策。Meder 等人（2017）提出个体在决策过程中同时表征一系列动态变化的价值评估可以作为一种灵活的选择机制，将经验获得的价值信息与价值的其他特征结合起来，从而在变化的环境中做出自适应的决策。为了更好的适应环境，个体可能依据外部环境或自身状态来灵活地调整对选项所赋予的价值，从而形成主观偏好。Ai 等人（2018）通过建立数学模型，将决策与记忆的动态提取过程相结合，证明了主观偏好变化与睡眠状态下相关记忆的巩固有关。更有价值的是，研究者们利用计算模型探究精神障碍（如创伤后应激障碍）和生理损伤（如基底核损伤）患者的学习机制，为其恢复正常功能的治疗提供有力证据（Brown et al., 2018; Zhu, Jiang, Scabini, Scabini, & Hsu, 2019）。这些研究对心理学以及临床医学领域的未来研究都有着重要的启示意义。

事实上，道德认知在日常生活和心理学中都占有举足轻重的地位。为阐明道德决策、道德判断和道德推理的认知过程和神经机制，将计算建模这一强大的手段运用于道德认知领域也是不可避免的。以下将回顾在道德认知及其他领域运用都比较广泛的计算模型——漂移扩散模型、效用模型、强化学习模型和分层高斯过筛器模型。

2.1 漂移扩散模型

漂移扩散模型（Drift Diffusion Models，DDM）最早由 Ratcliff（1978）开发，它把决策描述为一个连续的抽样过程，即带有噪声的信息从起点累积到对应于某一选项的边界或阈值（即标准），该选项被选中（Ratcliff & McKoon, 2008）。公式如下：

$$dy(t) = v(\Delta u) \cdot dt + \sigma \cdot dW$$

公式中$y(t)$是在时间 t 时积累的信息量；Δu是两个选项边界的差异；v是信息累积的速度（即漂移率）；σ是维纳过程dW的高斯噪声参数。此外，DDM 的参数还包括起始点偏移量、边界高度和非决策时间等。漂移率代表偏好强度，即个体倾向于某一选项的偏好越强烈，信息向该选项积累的速度就越快。每个选项均有一个边界，边界表示在做出反应之前必须积累的信息量。而积累过程是有噪声的，在任意时刻，信息可能指向两个边界中的一个，但更多的时候指向正确的边界。而非决策成分包括对刺激的编码（该刺激将驱动决策过程）和从刺激或记忆中提取构成决策基础的刺激的维度。DDM 可以将潜在的认知过程体现在模型不同的成分上。例如，信息积累的速度、边界高度和非决策过程的持续时间（Mormann, Malmaud, Huth, Koch, & Rangel, 2010; Lerche & Voss, 2019; Voss, Rothermund, & Voss, 2004）。而且 DDM 考虑了所有的行为数据，即正确反应和错误反应的反应时分布的形状和位置（Ratcliff, Smith, Brown, & Mckoon, 2016; Ratcliff, Thapar, & McKoon, 2004）。

DDM 最初适用于基本的知觉和记忆任务等的反应时研究，例如单项识别和联想识别任务（Ratcliff, 1978; Ratcliff, et al., 2004）、知觉任务（包括亮度、字母、注意定向等）等（Ratcliff, Thapar, & McKoon, 2003; Thapar, Ratcliff, & McKoon, 2003; Smith, Ratcliff, & Wolfgang, 2004）。近十至十五年间，DDM 在决策过程的心理和神经机制研究中变得越来越重要，包括感知觉决策、简单的运动决策和基于价值的决策等。Gold 和 Shadlen（2007）回顾基本的决策形成要素如何在大脑中实现，从而提出决策是一个权衡先验、证据和价值的过程，并描述了与关键决策要素（包括深思熟虑和情感认同）相对应的具体数学运算。他们也揭示了感知任务的速度——正确性权衡和简单运动任务的可变的反应时的一种基本机制——将变化的决策变量（随时间累积并存储证据）与固定标准进行比较的决策规则。此外，Krajbich, Armel 和 Rangel（2010）也用 DDM 对注视模式和选择之间的关系进行定量预测。结果发现，在 DDM 的简单扩展中，注视点参与价值整合过程，可以定量地解释注视点和选择之间的各种关系，以及一些相当大的选择偏差。而且 Krajbich 等人发现视觉注视过程与价值比较过程存在因果关系。即通过外源性操纵相对注视时间，个体可能对选择产生偏倚。Eikemo, Biele, Willoch, Thomsen 和 Leknes（2017）研究阿片类药物对健康人类基于价值的决策的调节时，用 DDM 拟合了正确率和反应时的数据，从而揭示两个决策子过程预期的双向药物效应。总之，DDM 可以描述个体如何使用先验、证据和价值来形成决策，揭示多种形式的决策（如知觉决策、简单的运动决策和基于价值的决策等）背后的一般原则。

2.2 效用模型

DDM 通常用于只有两个备选方案的实验任务（即二选一），且实验每个条件的试次数量要多，而效用模型（Utility Models）可以更好地解释有更多选项的情况。在经济学领域，效用函数用于衡量与一组商品和服务有关的偏好。效用常常与幸福感和满意度等有关，而这些难以直接观测。因此，经济学家利用效用函数来表征这些抽象的、不可直接测量的变量（Debreu,1954）。后来，效用函数被用于社会决策领域，它将可供选择的选项的价值传达给决策者，促使决策者选择价值（即效用）最大的选项。效用模型的简单公式如下（假设有两个选项）：

$$\Delta V = \mathrm{U}_A - \mathrm{U}_B$$

公式中U_A是选项 A 的效用；U_B是选项 B 的效用；ΔV是个体的主观价值。在每一个试次

中，被试对每个选项有不同的偏好，当且仅当被试更喜欢选项 A 而不是 B 时，A 的效用量才大于 B 。因此，当$\Delta V > 0$时个体才会选择选项 A。通常，之后会用 softmax 函数估计被试的选择概率。

在社会决策领域中，效用模型主要用于探讨社会偏好或道德偏好。研究者们将效用模型与功能磁共振成像相结合，研究社会价值的神经表征，以评估他们对自我和他人利益的分配(Liu et al., 2019; Qu, Météreau, Butera, Villeval, & Dreher, 2019; Zhong, Chark, Hsu, & Chew, 2016)。这种研究方法在一定程度上解读了代表自我和他人潜在利益的神经机制，对于理解社会决策至关重要。此外，Lopez-Persem, Rigoux, Bourgeois-Gironde, Daunizeau 和 Pessiglione (2017) 在不同任务中得到了相同的效用函数，并且对选择的预测准确性很高。这表明了可比较的效用函数不仅可以解释经济选择，而且可以解释不同的动机导向行为。值得注意的是，效用模型假设个体的偏好是固定的。因为如果根据价格或预算变化来改变人们的行为，将无**法确定行为变化在多大程度上是由于价格或预算变化还是偏好的改变所致。**

2.3 强化学习模型

上述的漂移扩散模型和效用模型被广泛应用于决策领域，而强化学习模型（Reinforcement Learning Models）则是解决决策中的不确定性问题以及各种学习问题的强大工具，包括与游戏相关的问题（如 Tesauro & Gerald, 1995）、自行车骑行问题（如 Randløv & Alstrøm, 1998）和机器人控制（如 Riedmiller, Gabel, Hafner, & Lange, 2009）等。许多不同的强化学习算法已经开发出来解决这些问题（Szepesvari, 2010; Sutton & Barto, 1998）。学习主体通过反复试验，形成刺激与结果关联来优化获得未来奖励的可能性，从而灵活地选择获得奖励的行为，这一过程被称为强化学习。强化学习的关键是预测误差，即预期事件和获得事件之间的差异，然后用于更新对环境中事件的信念（Sutton & Barto, 1998）。此外，强化学习模型中最典型和广泛使用的是 Rescorla-Wagner 模型，该模型通过预测误差信号表征学习，概念简单，计算效率高（Rescorla & Wagner, 1972）。Rescorla-Wagner 模型假设，在时间 k 时，大脑计算和更新行为变量 Q_k 的值如下：

$$Q_{k+1}=Q_k+\alpha\cdot\delta_k$$

公式中 α 是学习率；δ_k 是预测误差，在时间 k 收到的实际奖励与预期奖励之间的差值；Q_k 是当前的期望；Q_{k+1} 是个体对未来奖励的期望。强化学习系统的目标是学习一种行为策略，使个体选择的动作或行为获得最大累计奖赏值。

强化学习模型解释了基于行为和基于结果的价值表征之间的区别，将其与自动加工与控制加工联系起来，并精确地阐明了认知和情感机制对这两种类型的加工的贡献。一方面，基于模型的强化学习激活杏仁核、海马和眶额皮质等脑区（Andrews-Hanna, Reidler, Sepulcre, Poulin, & Buckner, 2010; Zsuga, Biro, Papp, Tajti, & Gesztelyi, 2016)。具体地，杏仁核与腹侧纹状体联合编码刺激（即预期结果之外的事件），而海马与腹侧纹状体联合编码上下文（即结果的偶然性）。此外，眶额皮层由海马和杏仁核驱动，将与奖励相关的信息整合到上下文框架中。因此，眶额皮层将提供关于预期奖励的信息，从而计算出奖励预期（Wallis, 2007）。另一方面，无模型强化学习也能够激活腹侧纹状体（Zsuga et al., 2016）。那么，眶额皮层提供的奖励预期信息反馈给无模型系统，基于腹侧纹状体的功能连通性，使腹侧纹状体可以将基于模型的奖励信息与无模型的奖赏预测误差相结合，计算腹侧纹状体发出的价值信号。所

以，基于模型的强化学习和无模型强化学习并非相互分离，而是具有功能连通性。

2.4 分层高斯过筛器模型

强化学习模型为简单的学习和决策行为及其神经基础的功能提供了强大的解释。但是在现实中，涉及许多刺激和动作的情况下，这些算法的学习效率低，不能及时捕捉人类学习的速度，而造成这种差异的一个原因是人类利用了现实世界任务中固有的结构来简化学习问题（Gershman & Niv, 2010）。所以，改进强化学习模型是不可避免的。Mathy, Daunizeau, Friston 和 Stephan（2011）受到 Behrens, Woolrich, Walton 和 Rushworth （2007）开创性工作的启发，提出一个分层高斯过筛器（Hierarchical Gaussian Filter, HGF）模型，用于在多种形式的不确定性（如环境波动和感知不确定性）下的个体学习。该模型包含了一个状态层次结构，这些状态在时间上演化为高斯随机游动（Gaussian random walks），每一个游动（除第一级水平外）的幅度大小由层次结构的下一个最高水平决定。水平之间的耦合由参数控制。这些参数编码了环境中关于高阶结构的先验信念，使模型能够解释学习中的个体差异包括个体间差异以及跨时间的个体差异。HGF 可以加工离散状态和连续状态，并且可以解释环境事件与感知状态之间的确定性和概率关系，能够推导出控制环境中突发事件的所有隐藏状态的后验期望的封闭式更新方程，使得 HGF 计算效率很高，能够实时学习。这些更新方程的形式类似于 Rescorla-Wagner 模型，为强化学习理论提供了一个贝叶斯类比。Rescorla-Wagner 模型的结构是：当前期望=前一期望+学习率×预测误差，HGF 的更新方程形式如下图 1：

$$\underbrace{\mu_2^{(k)}}_{\text{当前期望}} = \underbrace{\mu_2^{(k-1)}}_{\text{前一期望}} + \underbrace{\sigma_2^{(k)}}_{\text{学习率}} \underbrace{\left(\mu_1^{(k)} - s\left(\mu_2^{(k-1)}\right)\right)}_{\text{预测误差}}$$

$$\underbrace{\mu_3^{(k)}}_{\text{当前期望}} = \underbrace{\mu_3^{(k-1)}}_{\text{前一期望}} \underbrace{+\sigma_3^{(k)} \frac{\kappa}{2} \frac{e^{\kappa\mu_3^{(k-1)}+\omega}}{\sigma_2^{(k-1)} + e^{\kappa\mu_3^{(k-1)}+\omega}}}_{\text{学习率}} \times \underbrace{\left(\frac{\sigma_2^{(k)} + \left(\mu_2^{(k)} - \mu_2^{(k-1)}\right)^2}{\sigma_2^{(k-1)} + e^{\kappa\mu_3^{(k-1)}+\omega}} - 1\right)}_{\text{预测误差}}$$

图 1 分层高斯过筛器的更新方程与 Rescorla-Wagner 模型结构的对比。 $\mu^{(k-1)}$ 是前一后验概率；$\mu^{(k)}$ 是当前新的后验概率（具体参数参见 Mathy et al., 2011）。

Mathy 等人 (2014)进一步阐述 HGF 如何为加工感知中的不确定性提供一种通用的方法，将 HGF 的层次结构扩展到任意数量，探讨了如何通过更新方程中编码的变分自由能的最小化来适应各种形式的不确定性。总之，HGF 为理解正常和非正常学习提供了一个新的基础，它将强化学习置于一个通用的贝叶斯方法中，从而将其与概率论中的最优原则联系起来。它为解决行为者的感知不确定性提供了一个有原则的、灵活的、有效的同时又直观的框架。

HGF 是一种学习模型，它的特点是假定了个体进行社会学习时，形成关于他人印象的过程发生在多个认知层面上。在这里以两个认知层面：外显和内隐层面为例，外显可观测的层面是他人的具体行为，内隐（hidden）层面是观察者内心（或说头脑里）对他人的印象。HGF 可以计算给出外显层面的信息（即每次观察到他人的具体行为）如何推动内隐层面的表征的变化，即给出了一种“生成模型”。Siegel 等人（2018）选用 HGF 来探究个体道德推理的计算基础及其时间动态，就是因为它能解释内隐印象和外显观察到行为的关系，以说明外显行为观测如何推动印象形成。综上，借助实用的方法来开发人类认知的计算模型，这些模型基于可靠的概率原

理，可以解释日常思维、推理和学习的丰富性和复杂性。

3 计算模型在道德认知领域的运用

计算模型可以估计道德认知过程中内隐的、不可观测到的潜在成分（反映认知过程的参数）。研究者可以解释和预测这些潜在成分的具体认知加工过程，发展与完善道德认知的心理学理论。计算模型可以连接道德认知和道德神经科学，通过不同层面的计算模型，更加全面地解释和预测道德认知的神经机制。例如，研究者利用计算模型结合神经影像学，揭示心理学理论中潜在的、不能直接观察到的、与行为有关的神经活动过程和认知加工成分，如强化学习中的关键变量——奖赏预测误差（Sven, Pauli, Peter, & John, 2017）。本部分将回顾上面介绍的漂移扩散模型、效用模型、强化学习模型和分层高斯过筛器模型如何运用于道德认知领域。

3.1 计算模型在道德决策中的运用

人们在面对不同价值的选择时，并不总是依据利益最大化原则，选择价值更高的选项（Behrens, Hunt, & Rushworth, 2019; Crockett, Kurth-Nelson, Siegel, Dayan, & Dolan, 2014; Crockett et al., 2015）。有研究指出，人们考虑到他人的利益，而做出偏离自己利益最大化选择的程度与其道德行为呈正相关（Hutcherson et al., 2015; Yu et al., 2019）。

Hutcherson 等人（2015）让被试决定是否接受给自己和对家的分钱方案，探究人们的慷慨决策。在 DDM 中，每个试次的选择都基于动态变化的随机相对决策值信号——来估计相较于默认方案，对分配方案的预期。当随机相对决策值信号超过阈值时，被试会做出反应（如果是正值，接受分配方案；反之，则拒绝分配方案），反应时等于信息累积时间与非决策时间之和。结果发现，对他人的慷慨程度与自己的权重和启动阈值呈负相关，与漂移率呈正相关（Hutcherson et al., 2015; Konovalov & Krajbich, 2019）。此外，慷慨误差（错误地选择给予他人更多金钱）的比率明显高于自私误差（错误地选择保留更多金钱）的比率，这表明当个体获得的奖赏比别人获得的奖赏更有价值时，他/她的慷慨行为可能反映的是噪声干扰，而不是真正的亲社会偏好。在神经层面，个体在加工自己利益的过程中，腹内侧前额叶皮层和腹侧纹状体激活更强，而在加工他人利益的过程中，腹内侧前额叶皮层、右侧颞顶联合区和楔前叶激活更强。这表明加工自己利益和他人利益在大脑中是各自独立表征的。而且腹内侧前额叶皮层将关于自己利益和他人利益组合成一个整体值，并通过 DDM 的算法整合分配方案的总金额来做出选择。通过 DDM 对决策过程的随机相对决策值信号、漂移率、边界高度、起始点偏移量和非决策时间成分参数的拟合而推导和测试出，与自私决策相比，在做出慷慨决策前，与选项信息累积和价值计算相关脑区更活跃。这些研究结果揭示了道德价值表征背后的神经计算机制，并表明可能通过调节腹内侧前额叶皮层的道德价值表征来促进亲社会性。

Krajbich, Hare, Bartling, Morishima, 和 Fehr（2015）通过 DDM 发现社会决策（自私或慷慨）的速度和一致性可以通过从非社会决策（如食物选择）中得到的模型参数来预测，表明这两个领域的决策可能有着相同的加工模式。此外，对于社会决策是单一的比较过程还是双重过程（直觉的和深思熟虑的）问题，Chen 和 Krajbich（2018）提出归因于直觉的行为可以作为 DDM 过程的起点偏差，这种起点偏差类似于贝叶斯框架中的先验偏差。在独裁者博弈任务中，被试对如何在自己和对家之间分配金钱做出二元决策。结果发现，在时间压力下，

亲社会个体变得更亲社会，而在时间延迟下，亲社会个体数量变少。这些发现有助于统一关于社会决策认知加工过程的争论。

Crockett 等人（2014）让被试决定是否给自己和他人施加电击以换取利益（获得金钱数量随电击数量增加而增加），来探究人们的道德决策。Crockett 等人在效用模型中使用了——选项与默认选项之间的金钱差异和电击差异、损失厌恶参数和伤害厌恶参数，量化了被试给自己和他人带来的痛苦的相对价值。当伤害厌恶参数等于 0 时，决策者有最小的伤害厌恶，将会接受任何程度的电击来增加自己的收益；当伤害厌恶参数接近 1 时，决策者有最大的伤害厌恶，将会减少自己的收益来避免电击。之后，利用 softmax 函数将逐次试验的主观价值转化为选择概率。结果发现即使个体的决策完全是匿名的（未来不会受到不利的评判或惩罚），他们也更关心他人的痛苦，而不是自己的痛苦。而且这种对他人痛苦的关心与做出影响他人的决策时反应较慢有关，与道德决策过程中的深思熟虑一致。计算模型确定了这种亲社会倾向的精确边界，对于理解人类道德决策具有重要意义。

之后，Crockett 等人借助效用模型研究了道德决策中的生理和神经机制。结果发现，血清素水平的升高，增加了伤害厌恶和在决策时考虑的时间，而多巴胺水平的升高则恰恰相反（Crockett, et al., 2015）。血清素和多巴胺在调节道德行为中的这些独特作用，对社会功能障碍的潜在治疗具有重要意义。道德偏好较强的个体通过伤害他人获取利益时背侧纹状体激活较低，而外侧前额叶皮层编码了这种罪恶感（Crockett, Siegel, Kurth-Nelson, Dayan, & Dolan, 2017）。这表明伤害厌恶这种道德偏好可能会影响指导我们做出选择的价值观。值得注意的是，效用模型中的参数会随着不同的道德决策问题（如诚实、公平和慷慨）而有所变化（Gao et al., 2018; Hu et al., 2018; Sáez, Zhu, Set, Kayser, & Hsu, 2015; Strombach et al., 2015; Zhu et al., 2014）。

相较于传统研究方法，漂移扩散模型和效用模型展示了计算模型的价值，并为道德决策的本质提供了新的见解。它们都很好地解释和预测了自己利益和他人利益的权重对道德决策的影响。相较于非正式模型，漂移扩散模型和效用模型中的参数虽然会随着道德决策范式的变化而变化，但研究者们对其有统一的认识使得这些计算模型的解释力更强，更有利于它们应用于更多的领域中。

3.2 计算模型在道德判断中的运用

在社会中，由于某些行为会对其他个体产生影响，人们进而会判断这些行为对他人是有益或有害的。Hackel 和 Zaki（2018）采取改编的独裁者博弈实验范式，即在每轮游戏中，捐赠者（高财富和低财富）选择与接受者分享 20%或 50%的捐款，而接受者获得捐赠者分享的金额点数。接受者随机地与捐赠者（2 名高财富和 2 名低财富）配对，并反复选择与哪名捐赠者互动。因此，接受者同时了解每个捐赠者的慷慨程度（分享 20%捐款的慷慨程度为 0，分享 50%捐款的慷慨程度为 1）和奖励价值（20%或 50%×捐赠金额点数）。接下来，接受者完成一项互惠任务，与每位捐赠者分享金额点数作为回报。Hackel 和 Zaki 利用强化学习模型对接受者的互动选择进行了拟合，其中，奖赏预测误差反映了捐赠者的奖赏值和慷慨程度。例如，捐赠者先分享捐赠的 20%，后分享 50%，就会使接受者产生一个慷慨预测误差（即捐赠者表现得比接受者预期的更慷慨）。接受者对慷慨的捐赠者回报更多（Nowak & Sigmund, 2005），这是因为接受者对捐赠者进行了一个积极正面的道德判断，选择对其进行奖励，因此强化了自己的捐赠行为。在强化学习之后，人们不仅喜欢慷慨的社交伙伴，也喜欢那些提供大量物质奖励的人(Feldmanhall, Otto, & Phelps, 2018; Hackel, Doll, & Amodio, 2015; Hackel &

Zaki, 2018)。由此可以发现，道德判断是可以动态学习的，并引发了后续研究者在道德判断和亲社会行为的学习过程的深入探讨。

Yu 等人（2019）在效用模型和强化学习模型的基础上，提出一个以伤害厌恶为核心的计算模型，将道德决策、判断和推理的研究问题统一起来，为揭示道德认知的机制提供了独特的见解。在道德判断方面，个体在进行道德判断时，对行为者的责备程度与其选择伤害他人而产生的额外痛苦呈正相关，但与其选择伤害他人而产生的额外利益呈负相关。这表明，尽管个体会因损害他人利益责备行为者，但所获得的利益证明了部分伤害是合理的（Crockett, et al., 2010; Xie, Yu, Zhou, Sedikides, & Vohs, 2014）。总之，在道德判断中，获得利益和伤害他人对行为者受责备程度的影响相反。首先，人们认为伤害他人多于伤害自己获得利益，或者仅通过伤害他人获得利益，都会增加人们对不道德行为者的责备程度。其次，个体自己的伤害厌恶偏好调节获得利益和伤害他人对责备的影响，所以那些更不愿使他人痛苦的个体更关心伤害而不是收益，在判断行为者应该被责备或奖励时，会做出更极端的责备判断。综上，当行为者产生的负面结果影响他人时，基于伤害厌恶，会让判断者予以更多的惩罚，希望能够降低行为者伤害他人的行为。

除了伤害厌恶，道德判断也会涉及对不同规模和可能性的结果进行评估，例如电车困境中的获救人数和获救可能性。Shenhav 和 Greene（2010）让被试评估牺牲一条生命来拯救一个更大的群体的道德可接受性，这个群体的规模和不采取行动而死亡的可能性是不确定的，并基于简单的强化学习模型对数据进行拟合分析。结果发现，腹内侧前额叶皮层对生死道德判断中预期值的主观表征进行编码，而腹侧纹状体对预期道德价值特别敏感。同样，右侧前脑岛对死亡概率特别敏感。这表明，对影响他人生死攸关的复杂道德决策进行判断时依赖于适应更基本的、涉及物质奖励的自利决策的神经回路。Shenhav 和 Greene（2014）进一步利用基于模型的强化学习和无模型的强化学习对数据进行拟合分析，发现自动加工和控制加工对道德判断的影响之间的关键分离，且由不同的神经结构辅助。杏仁核激活反映了个体对有害的功利主义行为的厌恶和责备程度。在这种综合的道德判断中，腹内侧前额叶皮层优先参与相对功利主义和情感评价加工（Shenhav & Greene, 2014）。杏仁核和腹内侧前额叶皮层的功能连接随着任务中情绪输入所起的作用而变化，在纯功利主义判断中最低，在纯情绪判断中最高（Shenhav & Greene, 2010, 2014）。这些发现表明杏仁核对所判断的行为提供了情感评估，而腹内侧前额叶皮层则将这种信号与对预期结果的功利主义评估结合起来，得出经过深思熟虑的道德判断的结果。总之，研究者对道德认知的神经基础的探索发现，在道德判断过程中，大脑区域始终处于激活状态（Crockett et al., 2017; Shenhav & Greene, 2010）。进一步，计算模型可以精确地指定在道德判断过程中由大脑区域提供的计算。这促进了道德神经科学的发展，并加强了观察到的大脑和行为变化之间的联系。

3.3 计算模型在道德推理中的运用

道德推理是一个宽泛的概念，是个体指出影响他们进行道德评价的行为特征（如行为的结果和行为者的意图等）的过程，不一定是对善与恶的推理。一切通过社会学习去推断他人特征（如个体知觉和印象形成等）都可以看作道德推理（Feldmanhall, Dunsmoor, et al., 2018; Hackel et al., 2015; Joiner, Piva, Turrin, & Chang, 2017; Suzuki et al., 2012）。在社会互动中，推断他人的意图（intention）是形成道德印象的一个基本问题。而道德推理的一个基本挑战是人类如何了解他人的特征来预测自己的决策行为。研究表明，攻击者的道歉不仅会降低受害者的反应性攻击，还会改变攻击者对冒犯者的内隐态度（Beyens, Yu, Han, Zhang, & Zhou, 2015）。因此，某行为的道德性很大程度上取决于行为者的意图，对他人行为背后的意图进

行推断是道德判断和道德推理重要的环节。

Siegel 等人（2018）采用分层高斯过筛器来探究个体道德推理的计算基础及其时间动态。被试（正常大学生）预测并观察了两名行为者的一系列选择——是否对另一个人施加痛苦的电击以换取金钱，评估他们对行为者道德品质的印象以及不确定性。个体形成关于行为者道德品质的信念由概率分布表示，其中均值描述了每个试次后关于行为者的信念，并且方差描述了该信念的不确定性。信念随着时间的更新表征为高斯随机游动，其更新大小由表示信念波动的个体差异决定。结果表明，个体对不道德行为者的道德信念比对道德行为者的更具不确定性，并伴有更快的学习速度。这种机制可以使个体灵活地更新关于他人的信念。当最初的负面道德印象被证明不准确时，这种机制可以促进宽恕。

表 1　计算模型在道德认知研究中的应用总结

	道德认知		
	道德决策	道德判断	道德推理
漂移扩散模型	Chen & Krajbich, 2018 Hutcherson et al., 2015 Krajbich et al., 2015		
效用模型	Crockett et al., 2014, 2015, 2017 Gao et al., 2018 Hu et al., 2018 Sáez et al., 2015 Strombach et al., 2015 Yu et al., 2019 Zhu et al., 2014	Yu et al., 2019	Yu et al., 2019
强化学习模型	Yu et al., 2019	Hackel, et al., 2015 Hackel & Zaki, 2018 Shenhav & Greene, 2010, 2014 Yu et al., 2019	Hackel et al., 2015 Joiner et al., 2017 Suzuki et al., 2012 Yu et al., 2019
分层高斯过筛器模型			Siegel et al., 2018, 2019

注：Yu 等人（2019）在效用模型和强化学习模型的基础上提出一个以伤害厌恶为核心的计算模型，将道德决策、判断和推理的研究问题统一起来。

之后，Siegel 等人（2019）同样采用分层高斯过筛器研究男性服刑人员接触暴力对伤害学习的影响。结果发现接触暴力的个体形成了整体的主观社会印象，并将这些印象转化为社会决策，但会破坏其道德推理能力（认为道德行为者不值得信任，反而认为不道德行为者更

值得信任），从而导致更多的不道德行为。这是因为人们错误地把不好的特征归于好人会破坏现有的关系，阻碍建立新的关系（Johnson, Blumstein, Fowler, & Haselton, 2013）。因此，准确地推断他人道德品质的能力对健康的社会功能至关重要。从道德决策到道德推理是一个社会学习的过程，探究其认知和神经机制对于矫正服刑人员的认知、训练自闭症和抑郁症等精神障碍群体适应正常的社会功能等有重要意义。

Suzuki 等人（2012）利用强化学习模型，证明了个体模仿他人决策包括两个层次的学习信号。在模仿学习中，个体同时呈现两种不同的预测误差信号——模仿他人的奖赏预测误差和行为预测误差。个体模仿他人决策时，腹内侧前额叶皮层来模仿他人的特征以生成预测，并使用背内侧前额叶皮层和背外侧前额叶皮层来辅助行为变化以改进预测。Hackel 等人（2015）也利用强化学习模型揭示了个体在学习任务中通过反馈编码了奖赏和特征信息。除了特定的奖赏加工外，特征信息（如慷慨或自私等）通过反馈进行编码，并且在决策过程中，特征信息可以支配奖励信息。这两种学习方式都与腹侧纹状体的预测误差信号有关。对他人的印象也可以通过基于反馈的工具学习形成（Hackel et al., 2015）。简单举例阐述，某位同学与大家分享资源，可能不仅会收到回报，还被认为有慷慨、值得信任与合作等特质。于是她/他在其他情况下也会受到重视，比如更愿与其合作。此外，Joiner 等人（2017）讨论了自我参照和他人参照的奖赏预测误差，这些误差与多个大脑区域的激活有关（如纹状体、前扣带皮层、前额叶和颞顶联合区等），有效地使用强化学习模型来调节社会学习。计算模型的应用促进探索社会学习背后的神经机制，并增强了对道德推理的解释力。

4 不足与展望

道德行为和不道德行为在生活中普遍存在，但对其认知过程和神经机制的研究仍处于起步阶段。本文回顾了道德认知的三个维度（道德决策、判断和推理）以及几个在道德认知领域广泛运用的计算模型（漂移扩散模型、效用模型、强化学习模型和分层高斯过筛器模型），并梳理这些计算模型如何阐明道德心理的认知过程和神经机制。值得注意的是，漂移扩散模型、效用模型、强化学习模型和分层高斯过筛器模型与道德决策、判断和推理并不是一一对应的关系。计算模型更多地是与数据类型和实验设计相关，而心理过程上则可能没有这样的对应。例如，漂移扩散模型与强化学习模型结合使用，应用于道德认知的研究中，这可以作为研究者们将来研究的方向。相较于传统研究方法和非正式模型，计算模型准确地描述道德决策、判断和推理的认知过程，以及其潜在的神经关联。此外，研究者使用计算模型来研究道德领域的问题有助于解决关于道德认知中伤害的中心地位的争论（Schein & Gray, 2015, 2018）。

由于本文以道德认知领域为中心，所以无法详细讨论使用上述计算模型进行研究的其他领域，例如，资源分配（Konovalov et al., 2018）、精神障碍（Chen, Takahashi, Nakagawa, Inoue, & Kusumi, 2015; Rothkirch, Tonn, Köhler, & Sterzer, 2017）等。对这些领域的研究也明显受益于计算模型的使用。注意，这里所讨论的特定模型可能不会完全地适用于所有类型的社会行为，因此可能需要开发不同的计算方法。本文着重梳理了几个在道德认知领域广泛应用的计算模型以及它们如何应用于道德认知领域。所以，研究中也有其他能够解释道德认知问题的模型我们没顾及到，如多项加工树模型（预先规定了不同的过程如何作为实验输入和行为输出，主要用于对道德两难问题的研究; 刘媛媛, 丁一, 彭凯平, 胡传鹏，2019；Cameron, Payne, Sinnott-Armstrong, Scheffer, & Inzlicht, 2017；Gawronski, Conway, Armstrong, Friesdorf, & Hütter, 2018）和部分观测者马尔科夫决策过程模型（是贝叶斯模型的一种，主要用于探讨

社会情境下的信念学习；Khalvati et al., 2019）等。目前，没有一个单一的计算模型可以为道德认知提供一个明确和统一的机制。正如简单地为机器人提供一组“如果——那么”的规则来适应特定的情况非常困难，因为机器人可能发现自己处于无限多的情况中。从单一研究中得出的参数也不能作为适用于道德认知的各个组成部分的数字权重的最终结论。

前面的部分已经介绍了使用计算模型来研究道德认知的一些优点，这里强调与这种方法相关的潜在问题。首先，使用不同的模型来获得价值、信念或选择过程会存在一定的风险——模型的选择（而不是行为本身）决定了研究者研究的重点。例如，用于解释信念学习或偏好的模型（和任务）不同，至少驱动这些行为的过程中的一些差异反映了不同计算模型的使用。道德认知领域的进一步发展将会需要更统一的方法来对不同类型的认知进行建模。这个问题可以通过信任的相关研究来说明，信任主要表现为一个学习问题——由于信任他人使自己容易受到他人的背叛，人们必须解决潜在利益与至少三种其他担忧之间的冲突：损失厌恶、不公平厌恶和背叛厌恶（Bohnet & Zeckhauser, 2004）。很少有研究考察了在信任——不信任决策中这些厌恶背后的神经计算机制，这些机制可以用混合模型对这些不同的关注点分配权重来研究（Nave, Camerer, & McCullough, 2015）。

其次，虽然计算模型可以促进研究者对道德认知的理解和预测，但它们提供对潜在认知、学习和过程的看法有限。一些模型适合行为和大脑活动，这在很大程度上是因为它们能够灵活地适应许多不同模式的数据。因此，实证研究应该努力提供证据，证明模型的潜在参数实际上反映了可以通过实验干预选择性的改变（Hill et al., 2017）。最终，好的模型是那些能够构建关于驱动道德认知研究的模型，正如经典理论一样，但是现在有了一个更加定量和机械论的焦点。本质上，所有的模型都是错误的，但有些是有用的，可以为道德认知理论做出贡献。

最后，由于模型构建过程本身是比较多样和灵活的，因此如何能够保证计算模型不被滥用和误用也是非常重要的。Lee 等人（2019）提出了一种技术和实践方法，包括预注册模型、提供模型并在探索性模型开发后注册、对模型进行详细的评估和注册建模报告，使心理建模更加透明、可信、有效和稳定。构建适合计算建模的范例可能需要在现实世界的丰富性与方法论的严谨性之间进行权衡。识别一个对行为或大脑活动提供良好匹配的计算模型并不能保证所识别的模型是最好或最准确的模型（Mars, Shea, Kolling, & Rushworth, 2012）。此外，认知计算建模的一个重要且经常被忽视的方面是根据观测数据模拟候选模型（Palminteri, Wyart, & Koechlin, 2017）。尽管存在这些限制，但是计算模型有益于定量测量不依赖于自我报告的道德认知的个体差异，而自我报告在测量具有强烈社会赞许性成分的特征方面可能不太可靠。此外，模型参数可作为生物学和现象学的中间水平或认知表型，描述特定临床或亚临床和精神状态如何影响道德认知和行为，如抑郁症（Chen et al., 2015; Rothkirch et al., 2017）、精神分裂症（Valton, Romaniuk, Steele, Lawrie, & Seriès, 2017）和人格障碍（Tyrer, Reed, & Crawford, 2015）等。因此，使用计算模型不仅极大地促进我们对人类道德的理解，而且也逐渐地运用于计算精神病学和其他疾病，希望减少人类在疾病方面的痛苦。

5 结论

描述道德决策、判断和推理的计算模型代表了量化道德认知以及客观指导理解道德行为的认知过程的第一步。这些计算模型以数学方程的形式描述了道德选择的输入如何转化为输出。计算模型的优势在于它们提供了一种通用的数学语言，可以用来比较不同道德认知研究

的效果大小。随着越来越多的研究应用这些计算模型，研究者们将其汇总，可以上升到理论层面（如描述如何结合道德决策、判断和推理成分为道德认知领域完善某种理论或提出新的理论），也可以为临床领域提供经验和帮助（如计算精神病学）。目前，计算模型在道德认知领域的研究刚刚起步，相对少数的模型可以捕捉到道德认知的大部分方面，或者，人类道德的丰富性和复杂性可能无法归结为一组可管理的数学方程，这是有待研究者们解决的问题。

参考文献

参考消息. 中国"基因编辑婴儿"震惊世界！等待贺建奎的将是——. 2019–01–23 取自 http://ihl.cankaoxiaoxi.com/2018/1127/2359328.shtml

刘媛媛, 丁一, 彭凯平, 胡传鹏. (2019). 多项式加工树模型在社会心理学中的应用. *心理科学, 42*(2), 422–429.

Ai, S.Z., Yin, Y. L., Chen, Y., Wang, C., Sun, Y., Tang, X. D., ... Shi, J. (2018). Promoting subjective preferences in simple economic choices during nap. *eLife, 7*, e40583.

Andrews-Hanna, J. R., Reidler, J. S., Sepulcre, J., Poulin, R., & Buckner, R. L. (2010). Functional-anatomic fractionation of the brain's default network. *Neuron, 65*, 550–562.

Behrens, T. E., Hunt, L. T., & Rushworth, M. F. (2019). The computation of social behavior. *Science, 324*(5931), 1160–1164.

Behrens, T. E. J., Woolrich, M. W., Walton, M. E., & Rushworth, M. F. S. (2007). Learning the value of information in an uncertain world. *Nature Neuroscience, 10*(9), 1214–1221.

Beyens, U., Yu, H., Han, T., Zhang, L., & Zhou, X. (2015). The strength of a remorseful heart: Psychological and neural basis of how apology emolliates reactive aggression and promotes forgiveness. *Frontiers in psychology, 6*(1611), 1–16.

Bohnet, I., & Zeckhauser, R. (2004). Trust, risk and betrayal. *Journal of Economic Behavior & Organization, 55*(4), 467–484.

Brown, J. W. (2014). The tale of the neuroscientists and the computer: Why mechanistic theory matters. *Frontiers in neuroscience, 8*(349), 1–3.

Brown, V. M., Zhu, L., Wang, J. M., Frueh, B. C., King-Casas B., & Chiu, P. H. (2018). Associability-modulated loss learning is increased in posttraumatic stress disorder. *eLife, 7*, e30150.

Cameron, C. D., Payne, B. K., Sinnott-Armstrong, W., Scheffer, J. A., & Inzlicht, M. (2017). Implicit moral evaluations: A multinomial modeling approach. *Cognition, 158*, 224–241.

Charpentier, C. J., & O'Doherty, J. P. (2018). The application of computational models to social neuroscience: Promises and pitfalls. *Social Neuroscience, 13*(6), 637–647.

Chen, F., & Krajbich, I. (2018). Biased sequential sampling underlies the effects of time pressure and delay in social decision making. *Nature communications, 9*(3557), 1–10.

Chen, C., Takahashi, T., Nakagawa, S., Inoue, T., & Kusumi, I. (2015). Reinforcement learning in depression: A review of computational research. *Neuroscience & Biobehavioral Reviews, 55*, 247–267.

Cohn, A., Fehr, E., & Maréchal, M. A. (2014). Business culture and dishonesty in the banking industry. *Nature, 516*, 86–89.

Crockett, M. J., Clark, L., Hauser, M. D., & Robbins, T. W. (2010). Serotonin selectively influences moral judgment and behavior through effects on harm aversion. *Proceedings of the National Academy of Sciences, 107*(40), 17433–17438.

Crockett, M. J., Kurth-Nelson, Z., Siegel, J. Z., Dayan, P., & Dolan, R. J. (2014). Harm to others outweighs harm to self in moral decision making. *Proceedings of the National Academy of Sciences, 111*(48), 17320–17325.

Crockett, M. J., Siegel, J. Z., Kurth-Nelson, Z., Ousdal, O. T., Story, G., Frieband, C., ... Dolan, R. J. (2015). Dissociable effects of serotonin and dopamine on the valuation of harm in moral decision making. *Current Biology, 25*(14), 1852–1859.

Crockett, M. J., Siegel, J. Z., Kurth-Nelson, Z., Dayan, P., & Dolan, R. J. (2017). Moral transgressions corrupt neural representations of value. *Nature neuroscience, 20*(6), 879–885.

Debreu, G. (1954). Representation of a preference ordering by a numerical function. *Decision processes, 3*, 159–165.

Eikemo, M., Biele, G., Willoch, F., Thomsen, L., & Leknes, S. (2017). Opioid modulation of value-based decision-making in healthy humans. *Neuropsychopharmacology, 42*(9), 1833–1840.

Eisenegger, C., Naef, M., Snozzi, R., Heinrichs, M., & Fehr, E. (2010). Prejudice and truth about the effect of testosterone on human bargaining behaviour. *Nature, 463*(7279), 356–359.

Elqayam, S., Wilkinson, M. R., Thompson, V. A., Over, D. E., & Evans, J. S. B. (2017). Utilitarian moral judgment exclusively coheres with inference from is to ought. *Frontiers in psychology, 8*(1042),1–18.

Engelmann, J. B., & Fehr, E. (2016). The slippery slope of dishonesty. *Nature neuroscience, 19*(12), 1543–1544.

Feldmanhall, O., Dunsmoor, J. E., Tompary, A., Hunter, L. E., Todorov, A., & Phelps, E. A. (2018). Stimulus generalization as a mechanism for learning to trust. *Proceedings of the National Academy of Sciences of the United States of America, 115*(7), E1690–E1697.

Feldmanhall, O., Otto, A. R., & Phelps, E. A. (2018). Learning moral values: Another's desire to punish enhances one's own punitive behavior. *Journal of Experimental Psychology General, 147*(8),1211–1224.

Gächter, S., & Schulz, J. F. (2016). Intrinsic honesty and the prevalence of rule violations across societies. *Nature, 531*(7595), 496–499.

Gamer, M., Rill, H. G., Vossel, G., & Gödert, H. W. (2006). Psychophysiological and vocal measures in the detection of guilty knowledge. *International Journal of Psychophysiology, 60*(1), 76–87.

Gao, X., Yu, H., Sáez, I., Blue, P. R., Zhu, L., Hsu, M., & Zhou, X. (2018). Distinguishing neural correlates of context-dependent advantageous-and disadvantageous-inequity aversion. *Proceedings of the National Academy of Sciences, 115*(33), E7680–7689.

Garon, M., Lavallée, M. M., Estay, E. V., & Beauchamp, M. H. (2018). Visual encoding of social cues predicts sociomoral reasoning. *PloS one, 13*(7), e0201099.

Garrett, N., Lazzaro, S. C., Ariely, D., & Sharot, T. (2016). The brain adapts to dishonesty. *Nature Neuroscience*, *19*(12), 1727–1732.

Gawronski, B., Conway, P., Armstrong, J., Friesdorf, R., & Hütter, M. (2018). Effects of Incidental Emotions on Moral Dilemma Judgments: An Analysis Using the CNI Model. *Emotion, 18*(7), 989–1008.

Gershman, S. J., & Niv, Y. (2010). Learning latent structure: Carving nature at its joints. *Current Opinion in Neurobiology, 20*(2), 251–256.

Gold, J. I., & Shadlen, M. N. (2007). The neural basis of decision making. *Annual Review of Neuroscience, 30*, 535–574.

Gray, J. (1987). The economic approach to human behavior: Its prospects and limitations. In Radnitzky, G., Bernholz, P. (Eds.), *The Economic Method Applied Outside the Field of Economics* (pp. 33–49). New York: Paragon House Publishers.

Greene, J. D. (2007). Why are vmPFC patients more utilitarian? A dual-process theory of moral judgment explains. *Trends in Cognitive Sciences, 11*(8), 322–323.

Greene, J. (2014). *Moral tribes: Emotion, reason, and the gap between us and them*. Penguin.

Greene, J. D., & Paxton, J. M. (2009). Patterns of neural activity associated with honest and dishonest moral decisions. *Proceedings of the National Academy of Sciences*, *106*(30), 12506–12511.

Hackel, L. M., Doll, B. B., & Amodio, D. M. (2015). Instrumental learning of traits versus rewards: Dissociable neural correlates and effects on choice. *Nature Neuroscience*, *18*(9), 1233–1235.

Hackel, L. M., & Zaki, J. (2018). Propagation of economic inequality through reciprocity and reputation. *Psychological science, 29*(4), 604–613.

Hill, C. A., Suzuki, S., Polania, R., Moisa, M., O'Doherty, J. P., & Ruff, C. C. (2017). A causal account of the brain network computations underlying strategic social behavior. *Nature Neuroscience,20*(8),1142–1149.

Hu, Y., He, L., Zhang, L., Wölk, T., Dreher, J. C., & Weber, B. (2018). Spreading inequality: neural computations underlying paying-it-forward reciprocity. *Social cognitive and affective neuroscience, 13*(6), 578–589.

Hutcherson, C. A., Bushong, B., & Rangel, A. (2015). A neurocomputational model of altruistic choice and its implications. *Neuron, 87*(2), 451–462.

Jiang, J. F., Summerfield, C., & Egner, T. (2016). Visual prediction error spreads across object features in human visual cortex. *The Journal of Neuroscience: The Official Journal of the Society for Neuroscience, 36*(50), 12746–12763.

Joiner, J., Piva, M., Turrin, C., & Chang, S. W. C. (2017). Social learning through prediction error in the brain. *npj Science of Learning, 2*(1), 8, 1–9.

Johnson, D. D. P., Blumstein, D. T., Fowler, J. H., & Haselton, M. G. (2013). The evolution of error: Error management, cognitive

constraints, and adaptive decision-making biases. *Trends in Ecology & Evolution, 28*(8), 474–481.

Jordan, J. J., Sommers, R., Bloom, P., & Rand, D. G. (2017). Why do we hate hypocrites? Evidence for a theory of false signaling. *Psychological Science, 28*(3), 356–368.

Kamm, F. M. (2015). *The trolley problem mysteries*. Oxford University Press.

Khalvati, K., Park, S. A., Mirbagheri, S., Philippe, R., Sestito, M., Dreher, J. C., & Rao, R. P. (2019). Modeling other minds: Bayesian inference explains human choices in group decision-making. *Science Advances, 5*(11), eaax8783

Koenigs, M., Young, L., Adolphs, R., Tranel, D., Cushman, F., Hauser, M., & Damasio, A. (2007). Damage to the prefrontal cortex increases utilitarian moral judgements. *Nature, 446*(7138), 908–911.

Konovalov, A., Hu, J., & Ruff, C. C. (2018). Neurocomputational approaches to social behavior. *Current Opinion in Psychology, 24*, 41–47.

Konovalov, A., & Krajbich, I. (2019). Revealed indifference: Using response times to infer preferences. *Judgment and Decision Making, 14*(4), 381–394.

Krajbich, I., Armel, C., & Rangel, A. (2010). Visual fixations and the computation and comparison of value in simple choice. *Nature neuroscience, 13*(10), 1292–1298.

Krajbich, I., Hare, T., Bartling, B., Morishima, Y., & Fehr, E. (2015). A common mechanism underlying food choice and social decisions. *PLoS Computational Biology, 11*(10), e1004371.

Lee, M. D., Criss, A., Devezer, B., Donkin, C., Etz, A., Leite, F. P., . . . Vandekerckhove, J. (2019). Robust modeling in cognitive science. *Computational Brain & Behavior, 2*, 141–153.

Lerche, V., & Voss, A. (2019). Experimental validation of the diffusion model based on a slow response time paradigm. *Psychological Research*, 83(6),1194–1209.

Levine, E. E., Barasch, A., Rand, D. G., Berman, J. Z., & Small, D. A. (2018). Signaling emotion and reason in cooperation. *Journal of Experiment Psychology: General, 147*(5), 702–719.

Liu, Y., Li, S., Lin, W., Li, W., Yan, X., Wang, X., ... & Ma, Y. (2019). Oxytocin modulates social value representations in the amygdala. *Nature neuroscience, 22*(4), 633–644.

Lopez-Persem, A., Rigoux, L., Bourgeois-Gironde, S., Daunizeau, J., & Pessiglione, M. (2017). Choose, rate or squeeze: comparison of economic value functions elicited by different behavioral tasks. *PLoS Computational Biology, 13*(11), e1005848.

Mars, R. B., Shea, N. J., Kolling, N., & Rushworth, M. F. (2012). Model-based analyses: Promises, pitfalls, and example applications to the study of cognitive control. *Quarterly Journal of Experimental Psychology, 65*(2), 252–267.

Mathys, C., Daunizeau, J., Friston, K.J., & Stephan, K.E. (2011). A bayesian foundation for individual learning under uncertainty. *Frontiers in Human Neuroscience, 5*(39), 1–20.

Mathys, C. D., Lomakina, E. I., Daunizeau, J., Iglesias, S., Brodersen, K. H., Friston, K. J., & Stephan, K.E. (2014). Uncertainty in perception and the hierarchical gaussian filter. *Frontiers in Human Neuroscience, 8*(825), 1–24.

Meder, D., Kolling, N., Verhagen, L., Wittmann, M. K., Scholl, J., Madsen K. H., … Rushworth, M. F. S. (2017). Simultaneous representation of a spectrum of dynamically changing value estimates during decision making. *Nature Communications, 8*(1942), 1–11.

Mormann, M. M., Malmaud, J., Huth, A., Koch, C., & Rangel, A. (2010). The drift diffusion model can account for the accuracy and reaction time of value-based choices under high and low time pressure. *Judgment and Decision Making, 5*(6), 437–44.

Nave, G., Camerer, C., & McCullough, M. (2015). Does oxytocin increase trust in humans? A critical review of research. *Perspectives on Psychological Science, 10*(6), 772–789.

Nowak, M. A., & Sigmund, K. (2005). Evolution of indirect reciprocity. *Nature, 437*(7063), 1291–1298.

Palminteri, S., Wyart, V., & Koechlin, E. (2017). The importance of falsification in computational cognitive modeling. *Trends in cognitive sciences, 21*(6), 425–433.

Qu, C., Météreau, E., Butera, L., Villeval, M. C., & Dreher, J. C. (2019). Neurocomputational mechanisms at play when weighing concerns for extrinsic rewards, moral values, and social image. *PLoS biology, 17*(6), e3000283.

Randløv, J. & Alstrøm, P. (1998). *Learning to drive a bicycle using reinforcement learning and shaping*. Paper presented at the Proceedings of the Fifteenth International Conference on Machine Learning (USA), Madison, Wisconsin (pp. 463–471). The International Machine Learning Society.

Ratcliff, R. (1978). A theory of memory retrieval. *Psychological Review, 85*(2), 59–108.

Ratcliff, R., & McKoon, G. (2008). The diffusion decision model: Theory and data for two-choice decision tasks. *Neural computation, 20*(4), 873–922.

Ratcliff, R., Smith, P. L., Brown, S. D., & Mckoon, G. (2016). Diffusion decision model: Current issues and history. *Trends in Cognitive Sciences, 20*(4), 260–281.

Ratcliff, R., Thapar, A., & McKoon, G. (2004). A diffusion model analysis of the effects of aging on recognition memory. *Journal of Memory and Language, 50*(4), 408–424.

Ratcliff, R., Thapar, A., & Mckoon, G. (2003). A diffusion model analysis of the effects of aging on brightness discrimination. *Perception & Psychophysics, 65*(4), 523–535.

Rescorla, R. A., & Wagner, A. R. (1972). A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement. In Black, A. H., Prokasy, W. F. (Eds.), *Current Research and Theory* (pp. 64–99). New York: Appleton-Century-Crofts.

Riedmiller, M., Gabel, T., Hafner, R., & Lange, S. (2009). Reinforcement learning for robot soccer. *Autonomous Robots, 27*(1), 55–73.

Rothkirch, M., Tonn, J., Köhler, S., & Sterzer, P. (2017). Neural mechanisms of reinforcement learning in unmedicated patients with major depressive disorder. *Brain, 140*(4), 1147–1157.

Sáez, I., Zhu, L., Set, E., Kayser, A., & Hsu, M. (2015). Dopamine modulates egalitarian behavior in humans. *Current Biology, 25*(7), 912–919.

Schein, C., & Gray, K. (2015). The unifying moral dyad: Liberals and conservatives share the same harmbased moral template. *Personality and Social Psychology Bulletin, 41*(8), 1147–1163.

Schein, C., & Gray, K. (2018). The theory of dyadic morality: Reinventing moral judgment by redefining harm. *Personality and Social Psychology Review, 22*(1), 32–70.

Shenhav, A., & Greene, J. D. (2010). Moral judgments recruit domain-general valuation mechanisms to integrate representations of probability and magnitude. *Neuron, 67*(4), 667–677.

Shenhav, A., & Greene, J. D. (2014). Integrative moral judgment: Dissociating the roles of the amygdala and ventromedial prefrontal cortex. *Journal of Neuroscience, 34*(13), 4741–4749.

Siegel, J. Z., Estrada, S., Crockett, M. J., & Baskin-Sommers, A. (2019). Exposure to violence affects the development of moral impressions and trust behavior in incarcerated males. *Nature Communications, 10*(1942), 1–9.

Siegel, J. Z., Mathys, C., Rutledge, R. B., & Crockett, M. J. (2018). Beliefs about bad people are volatile. *Nature Human Behaviour, 2*(10), 750–256.

Smith, P. L., Ratcliff, R., & Wolfgang, B. J. (2004). Attention orienting and the time course of perceptual decisions: Response time distributions with masked and unmasked displays. *Vision Research, 44*(12), 1297–1320.

Strombach, T., Weber, B., Hangebrauk, Z., Kenning, P., Karipidis, I. I., Tobler, P. N., & Kalenscher, T. (2015). Social discounting involves modulation of neural value signals by temporoparietal junction. *Proceedings of the National Academy of Sciences, 112*(5), 1619–1624.

Sutton, R. S., & Barto, A. G. (1998). *Reinforcement learning: An introduction*. MIT press.

Suzuki, S., Harasawa, N., Ueno, K., Gardner, J. L., Ichinohe, N., Haruno, M., ... Nakahara, H. (2012). Learning to simulate others' decisions. *Neuron, 74*(6), 1125–1137.

Sven, C., Pauli, W. M., Peter, B., & John, O. (2017). Neural computations underlying inverse reinforcement learning in the human brain. *eLife, 6,* e29718.

Szepesvari, C. (2010). Algorithms for reinforcement learning. *In Synthesis Lectures on Artificial Intelligence and Machine Learning, 4.1*, 1–103.

Tesauro, & Gerald. (1995). Temporal difference learning and TD-Gammon. *Communications of the ACM, 38*(3), 58–68.

Thapar, A., Ratcliff, R., & Mckoon, G. (2003). A diffusion model analysis of the effects of aging on letter discrimination. *Psychol Aging, 18*(3), 415–429.

Tyrer, P., Reed, G. M., & Crawford, M. J. (2015). Classification, assessment, prevalence, and effect of personality disorder. *The Lancet, 385*(9969), 717–726.

Uhlmann, E. L., Pizarro, D. A., & Diermeier, D. (2015). A person-centered approach to moral judgment. *Perspectives on Psychological Science, 10*(1), 72–81.

Uhlmann, E. L., & Zhu, L. (2014). Acts, persons, and intuitions: Person-centered cues and gut reactions to harmless transgressions. *Social Psychological and Personality Science, 5*(3), 279–285.

Valton, V., Romaniuk, L., Steele, J. D., Lawrie, S., & Seriès, P. (2017). Comprehensive review: Computational modelling of schizophrenia. *Neuroscience & Biobehavioral Reviews, 83*, 631–646.

Voss, A., Rothermund, K., & Voss, J. (2004). Interpreting the parameters of the diffusion model: An empirical validation. *Memory & Cognition, 32*(7), 1206–1220.

Wallis, J. D. (2007). Orbitofrontal cortex and its contribution to decision-making. *Annual Review of Neuroscience, 30*, 31–56.

Xie, W., Yu, B., Zhou, X., Sedikides, C., & Vohs, K. D. (2014). Money, moral transgressions, and blame. *Journal of Consumer Psychology, 24*(3), 299–306.

Yu, H., Siegel, J.Z., Crockett, M.J. (2019). Modeling morality in 3-D: Decision-making, judgment, and inference. *Topics in Cognitive Science,11*(2), 409–432.

Zhong, S., Chark, R., Hsu, M., & Chew, S. H. (2016). Computational substrates of social norm enforcement by unaffected third parties. *NeuroImage, 129*, 95–104.

Zhu, L., Jenkins, A. C., Set, E., Scabini, D., Knight, R. T., Chiu, P. H., . . . & Hsu M. (2014). Damage to dorsolateral prefrontal cortex affects tradeoffs between honesty and self-interest. *Nature Neuroscience, 17* (10), 1319–1321.

Zhu, L. S., Jiang, Y. M., Scabini, D., Scabini, R. T., & Hsu, M. (2019). Patients with basal ganglia damage show preserved learning in an economic game. *Nature Communications, 10*(802), 1–10.

Zsuga, J., Biro, K., Papp, C., Tajti, G., & Gesztelyi, R. (2016). The “proactive” model of learning: Integrative framework for model-free and model-based reinforcement learning utilizing the associative learning-based proactive brain concept. *Behavioral neuroscience, 130*(1), 6–18.

The application of computational modelling in the studies of moral cognition

ZHANG Yinhua; LI Hong; WU Yin

(School of Psychology, Normal College, Shenzhen University; Shenzhen Key Laboratory of Affective and Social Cognitive Science, Shenzhen 518060, China)

Abstract: Moral cognition focuses on the processing of information underlying the moral behavior. Recently, researchers have begun to apply computational modelling to moral cognition as to explore how moral cognition is represented in the brain. However, the research on the computational modeling of moral cognition is still at its infancy. The application of computational modelling (the Drift Diffusion Models, Utility Models, Reinforcement Learning Models and Hierarchical Gaussian Filter) in the behavioral and physiological studies of moral cognition quantified the cognitive processes and neural mechanisms underlying moral decision-making, moral judgment, and moral inference. In addition, this new approach could help to understand antisocial behavior and mental disorders. Finally, the computational modeling needs to be improved and future research need to pay attention to the potential limitations.

Key words: moral cognition; computational modelling; moral decision; moral judgment; moral inference